

Internet Archive's S3 like server API

This document is intended for a user who is comfortable in the unix command line environment. It covers the technical details of using the archive's S3 like server API.

For info on S3:

<http://docs.amazonwebservices.com/AmazonS3/latest/index.html?Welcome.html>

For info on IA's item structure:

<http://www.archive.org/about/faqs.php>

(sorry!)

You can also look at an item's structure directly by clicking the HTTP link shown on a details page. ex: <http://archive.org/details/stats>

HINT: For best results use curl or libcurl version 7.19 or higher.

Available at: <http://curl.haxx.se/>

To get api keys for the archive's S3-Like API go to:

<http://www.archive.org/account/s3.php>

What the S3 API does:

- o Items (things with details pages) get mapped to S3 Buckets.
 - ie: <http://archive.org/details/stats> is also available as:
<http://s3.us.archive.org/stats>
or, per s3 dns bucket style:
<http://stats.s3.us.archive.org/>
 - Files within items are also available as S3 keys, ex:
<http://stats.s3.us.archive.org/downloadsPerDay.png>
- o Doing a PUT on the S3 endpoint will result in a new internet archive Item
- o Files may also be uploaded to a an Item in the same way keys are added, via S3 PUT.
 - When a file is added to an Item, it is staged in temporary storage and ingested via the Archive's content management system. This can take some time.

We strive to make the S3 API compatible enough with current client code.

Hopefully you can just global search and replace amazonaws.com with us.archive.org.

For the popular s3cmd (<http://s3tools.org/s3cmd>):

```
perl -pi -e 's/amazonaws.com/us.archive.org/g' S3/*
```

worked! Your mileage may vary of course.....

How this is different from normal S3:

- o DELETE bucket is not allowed.
- o Only the HTTP 1.1 REST interface is supported.
- o Archive is much more likely to issue 307 Location redirects than Amazon is.
 - Which means clients with good 100-Continue support are very nice to have
 - curl versions curl-7.19 and newer have excellent 100-continue support

- o ACLs are fake. permissions are: World readable, Item uploader writable.
- o We have a special Low Security request signing mode, which allows the request to be unsigned, and a password simply provided for the request.
- o POST and COPY are not implemented.
- o If you want to see the diagnostic log of an s3 endpoint append ?log to the url for the endpoint.
ex: `http://stats.ia310835.s3dns.us.archive.org:82/downloadsPerDay.png?log`
the log format may change at any time
- o Range requests are currently ignored.

There are special features of the archive s3 connector to support the easy uploading of items.

- o There is a combined upload and make item feature, just set the header:
`x-archive-auto-make-bucket:1`
- o An http header can specify metadata the ends up in `_meta.xml` at make bucket time.
 - o just add headers of form `x-archive-meta-$meta_name:$meta_value`
(or `x-amz-meta-$meta_name:$meta_value`)
 - o if you want multiple tags in `_meta.xml` you can put numbers in front:
`x-amz-meta01-$meta_name:$meta_value_a`
`x-amz-meta02-$meta_name:$meta_value_b`
 - o meta headers are sorted prior to tag generation when place in the xml
- o There is a cleartext password mode; Authorization header can be of form `'Authorization: LOW $accesskey:$secret'`

EXAMPLES:

- o these features combined allow single command document upload with curl:
 - For best results use curl-7.19 or newer

Text item (a PDF will be OCR'd):

```
curl --location --header 'x-amz-auto-make-bucket:1' \
  --header 'x-archive-meta01-collection:opensource' \
  --header 'x-archive-meta-mediatype:texts' \
  --header 'x-archive-meta-sponsor:Andrew W. Mellon Foundation' \
  --header 'x-archive-meta-language:eng' \
  --header "authorization: LOW $accesskey:$secret" \
  --upload-file /home/samuel/public_html/intro-to-k.pdf \
  http://s3.us.archive.org/sam-s3-test-08/demo-intro-to-k.pdf
```

Movie item (Will get flash video player on details page):

```
curl --location --header 'x-amz-auto-make-bucket:1' \
  --header 'x-archive-meta01-collection:opensource_movies' \
  --header 'x-archive-meta-mediatype:movies' \
  --header 'x-archive-meta-title:Ben plays piano.' \
  --header "authorization: LOW $accesskey:$secret" \
  --upload-file ben-2009-05-09.avi \
  http://s3.us.archive.org/ben-plays-piano/ben-plays-piano.avi
```

- o If you want to upload a file to an existing item:
 - requires curl-7.19 or newer

```
curl --location
      --header "authorization: LOW $accesskey:$secret" \
      --upload-file /home/samuel/public_html/intro-to-k.pdf \
      http://s3.us.archive.org/sam-s3-test-08/demo-intro-to-k.pdf
```

- o Although the s3 interface supports GET and HEAD, high performance downloads are achieved via the archive web infrastructure:

```
curl --location http://archive.org/download/sam-s3-test-08/demo-intro-to-k.pdf
```